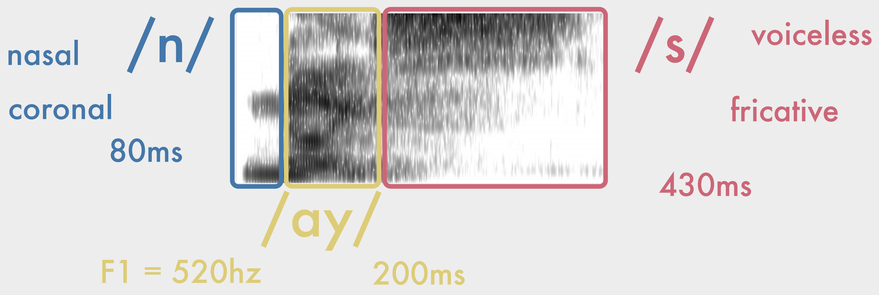
Choice of word frequency norms can dramatically affect inference

Choice of Word Frequency Norms can Dramatically Effect Inference.

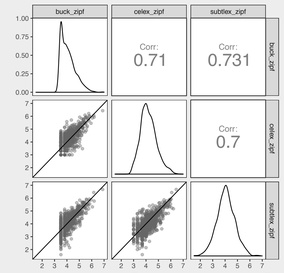
THE UNIVERSITY of EDINBURGH
School of Philosophy, Psychology and Language Sciences

Josef Fruehwald

Some factors influencing variation are observable, and others must be estimated. Different estimates may be correlated,



NICE
- valence
- lexical neighborhood density
- *frequency*



but are they interchangable?

Case study 1: TD Deletion

TD Deletion Data: Monomorphemes

west [wɛst] ~ [wɛs]
child [tʃaɪld] ~ [tʃaɪl]

Frequency Norms: Zipf Scaled

$\log_{10}$(frequency per million words) + 3

Model

td ~ zipfscore + (1 | Word) + (zipfscore | Speaker)

Regression Results:

Buckeye Corpus

6,691 Tokens

Frequency Norm	Estimated Effect	x Within Corpus
Within Corpus	−0.29	
Celex	−0.15	0.52
Subtlex	−0.10	0.34

Philadelphia Neighborhood Corpus

18,236 Tokens

Frequency Norm	Estimated Effect	x Within Corpus
Within Corpus	−0.599	
Celex	−0.006	0.01
Subtlex	−0.302	0.51

Discussion

The three different frequency norms result in very different estimated frequency effects. The within corpus frequency norm estimated a frequency effect twice to 100 times the size of the others.

Case study 2: /ay/ raising

/Data: /ay/ Raising from the PNC

right [ɹaɪt] ~ [ɹʌit]
nice [naɪs] ~ [nʌis]

18,608 F1 Estimates

Model:

F1 ~ decade * zipfscore + (decade | Word) + (zipfscore | Speaker)

Regression Results:

Within Corpus

	estimate	CI
intercept	0.68	(0.6, 0.76)
decade	−0.12	(−0.13, −0.10)
freq	−0.03	(−0.09, 0.04)
decade:freq	−0.006	(−0.01, 0.01)

Celex

	estimate	CI
intercept	0.64	(0.57, 0.71)
decade	−0.12	(−0.13, −0.10)
freq	−0.09	(−0.15, −0.01)
decade:freq	−0.001	(−0.01, 0.01)

Subtlex

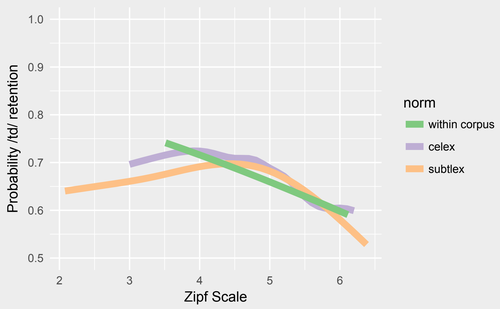
	estimate	CI
intercept	0.67	(0.6, 0.74)
decade	−0.12	(−0.13, −0.10)
freq	−0.05	(−0.12, 0.02)
decade:freq	−0.0003	(−0.01, 0.01)

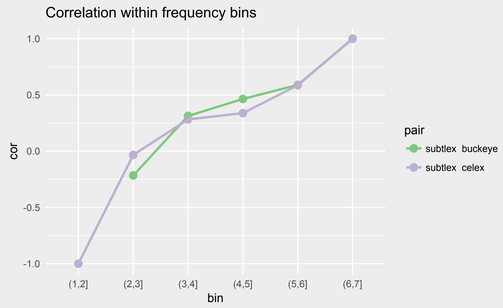
Discussion

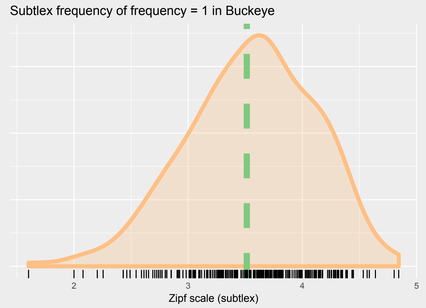
This time, the within-corpus frequency norm estimates the smallest frequency effect, but two of the norms don't have a reliable effect, while the remaining one does.

Why the differences?

The biggest difference between these norms is their estimates of low frequency words. *Recommendation: Use the norms with the best low frequency word estimates.*







References

Baayen, R. H., Piepenbrock, R. & Gulikers, L. (1995). The CELEX lexical database (Release 2, CD-ROM), LDC catalogue No.: LDC96L14, Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41(4), 977–90. https://doi.org/10.3758/BRM.41.4.977

Fruehwald, J. (2016). The early influence of phonology on a phonetic change. Language, 92(2), 376–410. https://doi.org/10.1353/lan.2016.0041

Hay, J. B., Pierrehumbert, J. B., Walker, A. J., & LaShell, P. (2015). Tracking word frequency effects through 130years of sound change. Cognition, 139, 83–91. https://doi.org/10.1016/j.cognition.2015.02.012

Labov, W., & Rosenfelder, I. (2011). The Philadelphia Neighborhood Corpus.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release). Columbus, OH. Retrieved from www.buckeyecorpus.osu.edu

Tamminga, M. (2014). Persistence in the Production of Linguistic Variation. University of Pennsylvania.

Van Heuven, W.J.B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. Quarterly Journal of Experimental Psychology, 67, 1176-1190